REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-01-

0164

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
19/01/01	Final Report	3/01/1997 - 06/30/2000

4. TITLE AND SUBTITLE
Large-Scale Optimization Methods with a Focus on Chemistry Problems

5a. CONTRACT NUMBER
F49620-97-1-0164

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Schnabel, Robert B.
Byrd, Richard. H.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Colorado at Boulder
Office of Contracts & Grants
3100 Marine St., Rm. 481, 572 UCB
Boulder, CO 80309

8. PERFORMING ORGANIZATION REPORT NUMBER
153-7606

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
801 N. Randolph St., Rm. 732
Arlington, VA 22203-1977

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of this research is to develop large-scale optimization methods for optimization problems that arise in molecular chemistry. The main applications that are being targeted are problems whose solution is of direct and immediate interest to the Air Force, such as finding the structure of proteins and polymers. The primary optimization problem being considered is the large-scale global optimization problem, via which protein structure can be determined. During this research period we have made major advances in the applicability of our methodology. The culmination of these advances has been our participation over the summer in the fourth CASP (Critical Assessment of Techniques for Protein Structure Prediction) competition, where we have attempted blind prediction of eight proteins of arbitrary complexity including mixed alpha helices and beta sheets, and of size ranging up to 242 amino acids. Enabling us to reach this stage have been advances in the past year in three key areas. The foremost of these is the ability to handle beta sheets in our algorithm, including the development of new biasing techniques. Second is improvements to the main portion of our global optimization method to enable it to better select the portions of the protein to work on in the small scale global optimizations. Third, in conjunction with our collaborators at Berkeley, is continued improvement of the ability of the energy function to accurately distinguish correct folds from misfolds, through the treatment of hydration.

15. SUBJECT TERMS
Global optimization, molecular chemistry, large-scale optimization, protein folding

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Schnabel, Robert B.
Unclassified	Unclassified	Unclassified		10	19b. TELEPHONE NUMBER (include area code) 303-492-7554

Final Technical Report

U.S. Air Force Grant F49620-97-1-0164
Large-Scale Optimization Methods with a Focus on Chemistry Problems

Richard Byrd, Robert Schnabel
Department of Computer Science
University of Colorado at Boulder
Boulder, CO 80309-0430

March 1, 1997 - June 30, 2000

2. Abstract

The objective of this research is to develop large-scale optimization methods for optimization problems that arise in molecular chemistry. The main applications that are being targeted are problems whose solution is of direct and immediate interest to the Air Force, such as finding the structure of proteins and polymers. The primary optimization problem being considered is the large-scale global optimization problem, via which protein structure can be determined. During this research period we have made major advances in the applicability of our methodology. The culmination of these advances has been our participation over the summer in the fourth CASP (Critical Assessment of Techniques for Protein Structure Prediction) competition, where we have attempted blind prediction of eight proteins of arbitrary complexity including mixed alpha helices and beta sheets, and of size ranging up to 242 amino acids. Enabling us to reach this stage have been advances in the past year in three key areas. The foremost of these is the ability to handle beta sheets in our algorithm, including the development of new biasing techniques. Second is improvements to the main portion of our global optimization method to enable it to better select the portions of the protein to work on in the small scale global optimizations. Third, in conjunction with our collaborators at Berkeley, is continued improvement of the ability of the energy function to accurately distinguish correct folds from misfolds, through the treatment of hydration.

2. Objectives

The overall objective of this research is to develop large-scale optimization methods for optimization problems that arise in molecular chemistry. The main applications that are being targeted are problems whose solution is of direct and immediate interest to the Air Force, such as finding the structure of polymers. The optimization approaches are being developed, however, in a manner that makes them applicable to a broad class of large-scale optimization problems. The main optimization problem that is being considered is the large-scale global optimization problem. This is the key problem that must be solved to determine the configuration of a molecule or macro-molecule once its potential energy function is known. Therefore the primary objective of this research is to develop efficient and effective large-scale global optimization methods for determining the structure of polymers and proteins. This includes testing of the methods on problems that constitute the state-of-the-art at the current time in optimization approaches to protein folding. One important component of this research is the development and implementation of new smoothing approaches within our global optimization approach. By transforming the energy landscape at intermediate stages of the solution process, these approaches appear to enable considerably more effective and efficient solutions of the global optimization problems. Another important component is working with chemists to determine potential energy functions that are effective in structural determination in an optimization context. Since the solution of large scale unconstrained optimization constitutes the major cost of the global optimization calculations, another research emphasis is finding ways to reduce the costs of the local optimizations in the context of the global optimization algorithm. In addition, due to the size and difficulty of the problems being solved, their solution requires the use of very powerful computers, generally parallel computers. Therefore development of efficient parallel large-scale global optimization methods is an objective of this research.

3. Status of Effort

We have continued to make considerable progress on a wide range of topics that are part of the development of effective, efficient optimization methods for protein folding problems. The foremost is that we have reached the stage in our research where we can attempt to predict the structure of arbitrarily

structured proteins of the size (100-250 amino acids) that is at the forefront of research in this field. To accomplish this we conducted key research this year in being able to handle beta sheets, the second and more difficult type of secondary structure within proteins. (The other type of structure, alpha helices, was our emphasis in the last year or two.) The main aspects of this research were the development of biasing functions that allow us to form beta sheets from predictions of secondary structure, and the development of methods to choose from among the many possible orientations of beta sheets. The second key aspect was the refinment of our methods in phase two of our algorithm, the main global optimization phase, to select both which proteins to work on and the subset of parameters for the key small scale global optimization step. The culmination of this research has been our entry into the CASP4 competition this summer, by participating in the blind prediction of eight of the proteins in that competition. In addition, we have continued our collaborations with Jorge Nocedal at Northwestern University on the development of a robust algorithm for nonlinearly constrained optimization. Finally, we have begun new work on efficient tensor methods for large systems of nonlinear equations.

4. Accomplishments/New Findings

Our main activity in this research period has continued to be the development and testing of techniques for solving global optimization problems for determining the structure of proteins and polymers. The problem is to find the lowest energy configuration of a protein or other polymer. This problem is a global optimization problem because it has a huge number of local minimizers. In addition, locating the lowest (global) minimizer is very difficult. For proteins, the solution of this problem would represent a solution to the well-known protein-folding problem. For the Air Force, one of the primary applications of this problem is in the development of new materials. For example, there is extensive work on this problem at the Air Force's Wright Laboratory for this reason.

In previous research periods, we have developed a stochastic/perturbation approach for solving global optimization problems from molecular chemistry. There are four keys to this approach. The first is a large scale global optimization methodology that performs small-scale global optimizations with only a small number of parameters variable and the remaining parameters temporarily fixed, followed by local minimizations with all parameters varying, at each stage of the global optimization procedure. The second is the incorporation of a new, efficient approach towards smoothing the objective function in the global optimization framework. Initially these have been the backbone of the approach. More recently two other aspects have become key to doing work on realistic protein targets. One of these is the incorporation of predictions from secondary structure prediction methods in the initial phase of our algorithm, to produce starting configurations with reasonably good secondary structure. The final one is work with our chemistry partners to use the mismatch between simulation and experimations to continue to refine the mathematical energy model upon which our global optimization approach relies.

Until this year, our research had been applied primarily to molecular clusters, and to small, alpha-helical proteins. We had developed good biasing methods for creating predicted alpha-helical secondary structure at the start of our method, and, in the last research period, had shown that our global optimization approach could do a reasonably good job of predicting the full tertiary structure of several helical proteins of about 70 amino acids. We had also demonstrated the effectiveness of our smoothing approach.

In the course of this year, we evolved our approach to be able to handle proteins with arbitrary structure. The crux of this issue, for us and other groups doing related research, is the ability to handle beta-sheets. Beta-sheets are the other main type of secondary structure in proteins. However they are far

less local than alpha-helices. While alpha-helices are continuous, beta-sheets are formed by contiguous strands that can be arbitrarily far apart. Secondary structure prediction programs can predict the strands with good accuracy, but they do not predict which strands are bonded together, nor the parallel or anti-parallel orientation of those bonds.

One main component of our research was the development of a biasing function that, given predictions of which amino-acids are bonded together to form the beta-sheet, influences the protein to form these bonds. Biasing functions are simply penalty functions from optimization that are added to the energy function. We constructed a function that is a combination of a sigmoid at low distances and a linear function at higher distances, to balance the need to form the bonds but not to overly bias. Along with this, we built upon existing software from the bio-chemistry community to construct techniques to predict the several most likely combinations of the predicted beta strands into beta sheets. The tests we describe later in this section, as well as earlier, simpler tests, clearly demonstrated the viability of these new approaches.

Simultaneously with the beta sheet research, we continued to refine the heart of the global optimization algorithm. As the protein sizes increase, the selection of roughly 5-6 dihedral angles (out of 200-500) at each stage to be the parameters in the small scale global optimization becomes even more crucial. And, not only should these be the parameters whose variation can lead to improvements in the tertiary structure, but they also must be a set that can be varied without destroying good secondary structure. We have developed new approaches to selecting these parameters that emphasize selecting from portions of the protein that are not parts of the regions of secondary structure. We also have begun to develop ways to selecting a set of angles from turns connecting beta sheets so as not to destroy the beta sheet.

Equally important is the ability to determine structural different proteins to work upon. This issue generalizes to any global optimization problem where we want to spread our effort over the search space. During this research period we formulated and implemented a clustering technique that takes all the currently active configurations and groups them into clusters of similarly shaped configurations. Then our algorithm only proceeds with one configuration from any given cluster at once. The clustering also is used to determine when to stop our algorithms, based upon whether the number of clusters still is growing, or not.

The culmination of our research this year was our participation in the fourth Critical Assessment of Techniques for Protein Structure Prediction (CASP4) competition in summer 2000. This competition, held every two years in the bio-chemistry community, invites any interested groups to blindly predict the structure of proteins that are about to be experimentally analyzed. About 170 groups entered this year. These groups utilize many different approaches, most based upon comparison to the structures of known proteins. Our approach, which only uses secondary strucuture prediction but not sequence matching, is at the pure end of the spectrum and is particularly important for predicting "new folds" that do not closely match known proteins. We spent the summer predicting eight proteins, with sizes ranging from 56 to 242 amino acids. Several of these were helical but the majority were mixed alpha-beta proteins. The results of the competition will only be known in December but it was clear that our methods were allowing us to form predicted beta structure and what appeared to be reasonable tertiary structures. (Note: since this report is late, we know the results now and our group did quite well, with results in the top quartile of all groups for the 3 proteins that were in the top 15% of difficulty, and the best result of all groups for the hardest protein we attempted with 242 amino-acids.)

During this research period we also began work on new tensor methods for very large scale of non-linear equations. We have developed a new approach for iteratively solving the tensor model that avoids the cost of a second backsolve that the previous approach had. We have also developed a new curvilinear line search for tensor methods that eliminates the need to use the tensor and Newton direction separately and which produces monotonic descent on the tensor model.

5. Personnel Supported

Personnel directly supported by this research grant:

Robert Schnabel, Principal Investigator

Richard Byrd, Co-Principal Investigator

Elizabeth Eskow, Professional Research Associate

Ms. Eskow is a long-time research staff member who has been mainly supported by NSF and our university in conjunction with computer science infrastructure grants, but is also partially supported by this grant. She plays an extensive role in all aspects of this project. In this research period she has been the main person conducting the computational testing for the CASP4 competition, and revising the code to deal with beta sheets. She also has played a large role in developing and testing the new heuristics in the global optimization method. Ms. Eskow also provides assistance and guidance to many of our graduate students and visitors, and interacts extensively with our research partners at Lawrence Berkeley Laboratory.

Brett Bader, Graduate Student

Mr. Bader joined our group as a new Ph.D. student in fall 1998. With a B.S. and M.S. from MIT in Chemical Engineering and several years work experience in this area that gave him extensive experience in optimization, he is ideally suited for our research in optimization methods for molecular chemistry problems. During the last year he contributed extensively to research about biasing functions for beta sheets, and also began thesis research on tensor methods for large systems of nonlinear equations.

6. Publications

Submitted but not yet accepted:

R.H. Byrd and H. Khalfan, "Analysis of a symmetric rank-one trust region method for constrained minimization", submitted for journal publication.

H. Khalfan, R.H. Byrd, and R. Schnabel, "Retaining convergence properties of trust region methods without extra gradient evaluations" submitted for journal publication.

A. Azmi, R. Byrd, E. Eskow and R. Schnabel, "New smoothing techniques for global optimization in solving for protein conformation", submitted for publication.

R. Byrd, J. Nocedal and R. Waltz, "Feasible Interior Methods Using Slacks for Nonlinear Optimization," submitted for journal publication.

Accepted but not yet published:

A. Bouaricha and R. Schnabel, "Tensor methods for large sparse nonlinear least squares problems", to appear in *SIAM Journal on Scientific Computing.*

D. Feng and R. Schnabel, "Local convergence analysis of tensor and SQP methods for singular constrained optimization", to appear in *SIAM Journal on Optimization.*

Published

Y. Xie and R. Byrd, "Practical update criteria for reduced Hessian successive quadratic programming algorithms," *SIAM Journal on Optimization.* 9, 1999, pp. 578-604.

R.H. Byrd, M. E. Hribar and J. Nocedal, "An Interior Point Algorithm for Large Scale Nonlinear Programming," *SIAM Journal on Optimization* 9, 1999, pp. 877-900.

C. Shao, R. Byrd, E. Eskow and R. Schnabel, "Global optimization for molecular clusters using a new smoothing approach", *Journal of Global Optimization* 16, 2000, pp. 167-196.

A. Azmi, R. Byrd, E. Eskow, R. Schnabel, S. Crivelli, T. Philip and T. Head-Gordon, "Predicting Protein Tertiary Structure Using a Global Optimization Algorithm with Smoothing", *Optimization in Computational Chemistry and Molecular Biology: Nonconvex Optimization and Its Applications,* C.A. Floudas and P.M. Pardalos, eds., Kluwer Academic Publishers, pp. 1-18, 2000.

R.H. Byrd, J.C. Gilbert and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming* 89, 2000, pp. 149-185.

7. Interactions / Transitions

7a. Meetings, Conferences, Seminars:

R. Byrd, "Step computation in a trust region interior point method," First Workshop on Nonlinear Optimization, University of Coimbra, Portugal, October 18, 1999.

R. Byrd, "False Convergence in Optimization Algorithms", International Conference on

Mathematical Programming, Atlanta, August 2000.

E. Eskow, "A Stochastic_Perturbation Global Optimization Approach to Protein Structure Prediction", International Conference on Mathematical Programming, Atlanta, August 2000.

R. Schnabel, "Optimization applied to Protein Folding: Interrelationships and Recent Progress", International Conference on Mathematical Programming, Atlanta, August 2000.

R. Schnabel, "Protein Structure Prediction by Global Optimization Utilizing Secondary Structure Prediction", SIAM Conference on Scientific Computing, Washington D.C., Sept. 2000.

7b. Consultation to other Laboratories and Agencies:

none

7c. Transitions:

The current research on global optimization methods for molecular configuration problems is at too early of a stage for transition to commercial or applied use. However we have had important interactions on this research outside of the optimization community that may help prepare for such transitions. We have a very active collaboration with Dr. Teresa Head-Gordon, a bio-chemist who has recently moved from the DOE Lawrence Berkeley Laboratory to a tenure-track faculty position at Berkeley. Our interaction with Head-Gordon's group has been very close, including joint development of energy models and optimization approaches and joint participation in the CASP4 conference. This collaboration includes almost daily phone or email contact and visits by Head-Gordon to Colorado and by our group to Berkeley.

We have also had important interactions with the wide protein folding community through our participation in several conferences, including the CASP4 conference in Monterey in Dec. 2000. Our interaction with the well-known protein folding research at Cornell University is aided by Dr. Schnabel's chairing of an annual evaluation committee at Cornell for a NIH program where Dr. H. Scheraga is one of the three principal investigators.

More generally, the research software that our group has produced is used very extensively in industrial and commercial work throughout the world. The most prominent example is our UNCMIN unconstrained optimization package that has been distributed to hundreds of sites and is included in several books (besides our own). It is also the basis of the unconstrained optimization software in the IMSL library. UNCMIN has been used in very many real applications but of course we are only aware of a tiny fraction of its use. During this past year we have continued to fulfill at least 1-2 requests a month for the UNCMIN software; the re-publication of our book, "Numerical Methods for Unconstrained Optimization and Nonlinear Equations" (J.E. Dennis Jr. and R. B. Schnabel) by SIAM in Feb. 1996 contributes to the ongoing demand. Another important example is our software for the modified Cholesky factorization which we know has been adopted in a variety of applications, including as the preferred option in the widely distributed LANCELOT optimization software package. A third example is the ODRPACK software for nonlinear regression that is extensively used in data fitting throughout the world. A fourth set of examples is our more recently released software for solving nonlinear equations and unconstrained optimization problems by tensor methods, for which we have received a good number of requests from industry in the past year and several reports that it enabled the user to solve problems more effectively than before. This software was produced in part from research supported by AFOSR. In general, each of

these packages is used in both industrial and academic settings, but don't know the full extent of its application.

8. New Discoveries, Inventions, or Patent Disclosures

none during this period

9. Honors / Awards

none during this period